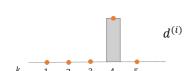
Introduction to Machine Learning Information Theory for Machine Learning





Learning goals

- Minimizing KL = maximizing log-likelihood
- Minimizing KL = minimizing cross-entropy
- Minimizing CE between modeled and observed probabilities = log-loss minimization

KL VS MAXIMUM LIKELIHOOD

Minimizing KL between the true distribution p(x) and approximating model $q(x|\theta)$ is equivalent to maximizing the log-likelihood.

$$egin{aligned} D_{\mathcal{KL}}(p\|q_{m{ heta}}) &= \mathbb{E}_{X \sim p} \left[\log rac{p(x)}{q(x|m{ heta})}
ight] \ &= \mathbb{E}_{X \sim p} \log p(x) - \mathbb{E}_{X \sim p} \log q(x|m{ heta}) \end{aligned}$$

as first term above does not depend on θ . Therefore,

$$rg \min_{m{ heta}} D_{\mathit{KL}}(m{p} \| q_{m{ heta}}) = rg \min_{m{ heta}} - \mathbb{E}_{X \sim m{p}} \log q(x | m{ heta})$$

$$= rg \max_{m{ heta}} \mathbb{E}_{X \sim m{p}} \log q(x | m{ heta})$$

For a finite dataset of n samples from p, this is approximated as

$$rg \max_{m{ heta}} \mathbb{E}_{X \sim p} \log q(x|m{ heta}) pprox rg \max_{m{ heta}} rac{1}{n} \sum_{i=1}^n \log q(\mathbf{x}^{(i)}|m{ heta})$$
 .

This also directly implies an equivalence to risk minimization!



KL VS CROSS-ENTROPY

From this here we can see much more:

$$\mathop{\arg\min}_{\boldsymbol{\theta}} D_{\mathit{KL}}(\boldsymbol{p} \| q_{\boldsymbol{\theta}}) = \mathop{\arg\min}_{\boldsymbol{\theta}} - \mathbb{E}_{\boldsymbol{X} \sim \boldsymbol{p}} \log q(\boldsymbol{x} | \boldsymbol{\theta}) = \mathop{\arg\min}_{\boldsymbol{\theta}} H(\boldsymbol{p} \| q_{\boldsymbol{\theta}})$$

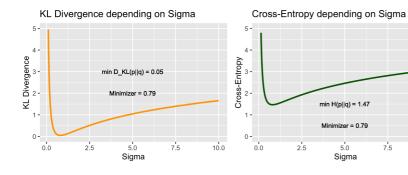
- So minimizing KL is the same as minimizing CE, is the same as maximum likelihood!
- We could now motivate CE as the "relevant" term that you have to minimize when you minimize KL after you drop $\mathbb{E}_p \log p(x)$, which is simply the neg. entropy H(p)!
- Or we could say: CE between p and q is simply the expected negative log-likelihood of q, when our data comes from p!



KL VS CROSS-ENTROPY EXAMPLE

Let
$$p(x) = N(0, 1)$$
 and $q(x) = LP(0, \sigma)$ and consider again $\arg\min_{\theta} D_{KL}(p||q_{\theta}) = \arg\min_{\theta} -\mathbb{E}_{X \sim p} \log q(x|\theta) = \arg\min_{\theta} H(p||q_{\theta})$



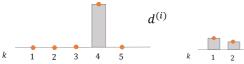


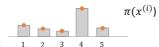
7.5

10.0

CROSS-ENTROPY VS. LOG-LOSS

- Consider a multi-class classification task with dataset $\mathcal{D} = ((\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)})).$
- For g classes, each $y^{(i)}$ can be one-hot-encoded as a vector $d^{(i)}$ of length g. $d^{(i)}$ can be interpreted as a categorical distribution which puts all its probability mass on the true label $y^{(i)}$ of $\mathbf{x}^{(i)}$.
- $\pi(\mathbf{x}^{(i)}|\theta)$ is the probability output vector of the model, and also a categorical distribution over the classes.







CROSS-ENTROPY VS. LOG-LOSS

To train the model, we minimize KL between $d^{(i)}$ and $\pi(\mathbf{x}^{(i)}|\boldsymbol{\theta})$:

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n D_{\mathit{KL}}(\boldsymbol{d}^{(i)} \| \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta})) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n H(\boldsymbol{d}^{(i)} \| \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta}))$$

We see that this is equivalent to log-loss risk minimization!

$$R = \sum_{i=1}^{n} H(d^{(i)} || \pi_k(\mathbf{x}^{(i)} | \boldsymbol{\theta}))$$

$$= \sum_{i=1}^{n} \left(-\sum_{k} d_k^{(i)} \log \pi_k(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \right)$$

$$= \sum_{i=1}^{n} \underbrace{\left(-\sum_{k=1}^{g} [y^{(i)} = k] \log \pi_k(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \right)}_{\log \log s}$$

$$= \sum_{i=1}^{n} (-\log \pi_{y^{(i)}}(\mathbf{x}^{(i)} | \boldsymbol{\theta}))$$



CROSS-ENTROPY VS. BERNOULLI LOSS

For completeness sake:

Let us use the Bernoulli loss for binary classification:

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x}))$$

If p represents a Ber(y) distribution (so deterministic, where the true label receives probability mass 1) and we also interpret $\pi(\mathbf{x})$ as a Bernoulli distribution Ber($\pi(\mathbf{x})$), the Bernoulli loss $L(y, \pi(\mathbf{x}))$ is the cross-entropy $H(p||\pi(\mathbf{x}))$.



ENTROPY AS PREDICTION LOSS

Assume log-loss for a situation where you only model with a constant probability vector π . We know the optimal model under that loss:

$$\pi_k = \frac{n_k}{n} = \frac{\sum_{i=1}^n [y^{(i)} = k]}{n}$$

What is the (average) risk of that minimal constant model?

$$\mathcal{R} = \frac{1}{n} \sum_{i=1}^{n} \left(-\sum_{k=1}^{g} [y^{(i)} = k] \log \pi_k \right) = -\frac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{n} [y^{(i)} = k] \log \pi_k$$
$$= -\sum_{k=1}^{g} \frac{n_k}{n} \log \pi_k = -\sum_{k=1}^{g} \pi_k \log \pi_k = H(\pi)$$

So entropy is the (average) risk of the optimal "observed class frequency" model under log-loss!

